User Repurchase Prediction

Sean Matthews

github/sean-io/market-basket-analysis

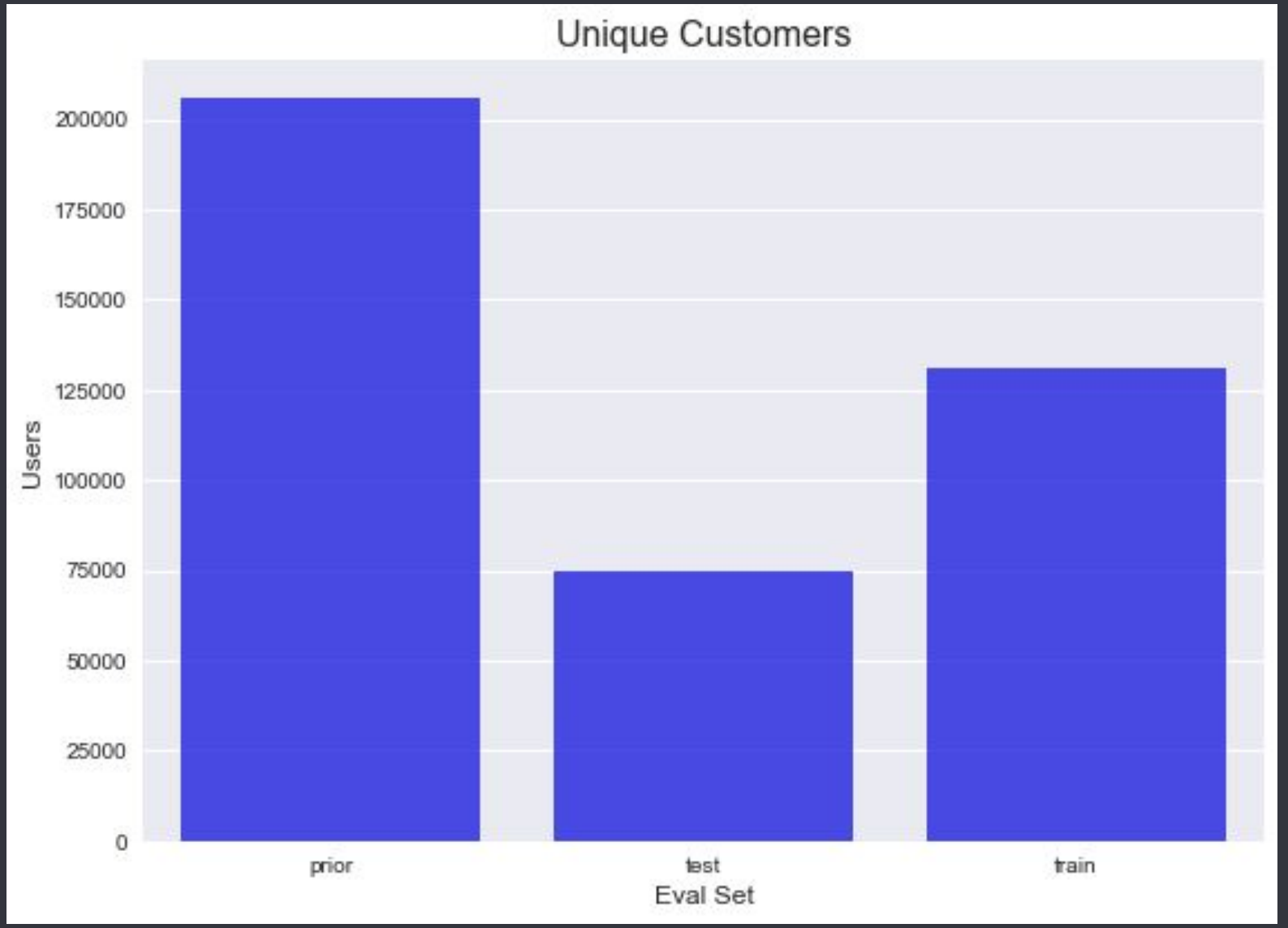
Objective & Context

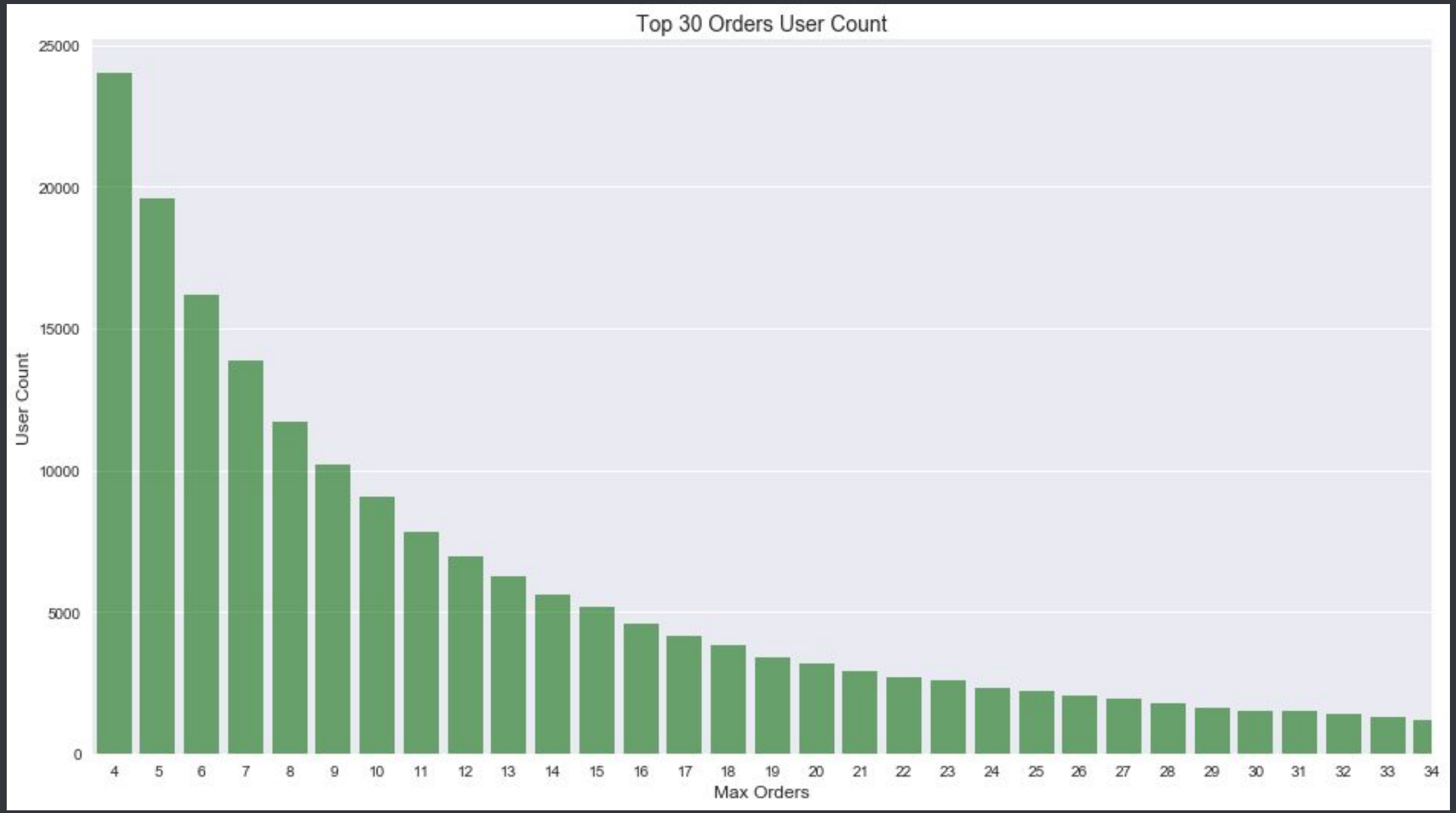
Predict which previously purchased products a user will order next.

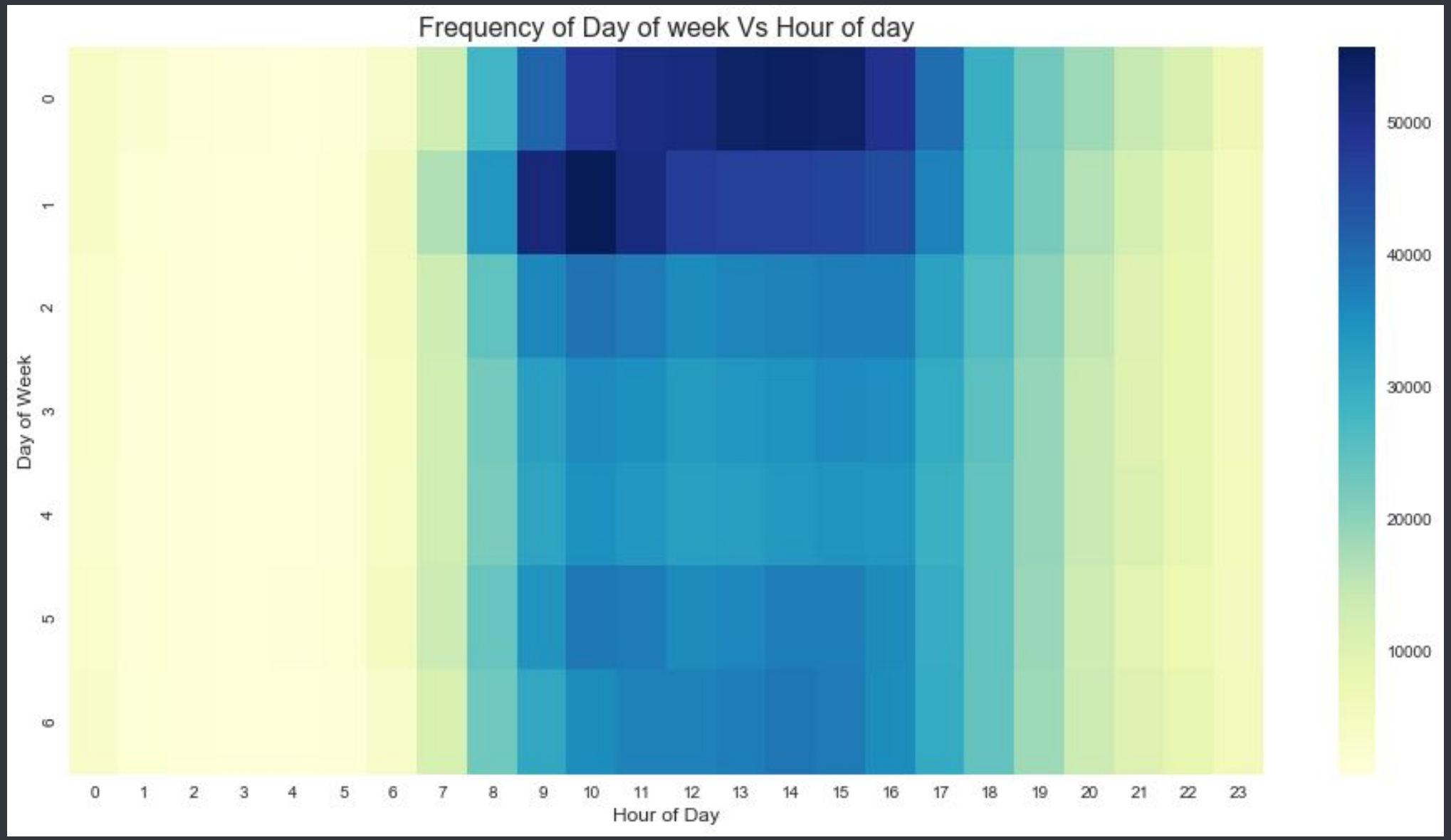
Kaggle Instacart Challenge Data

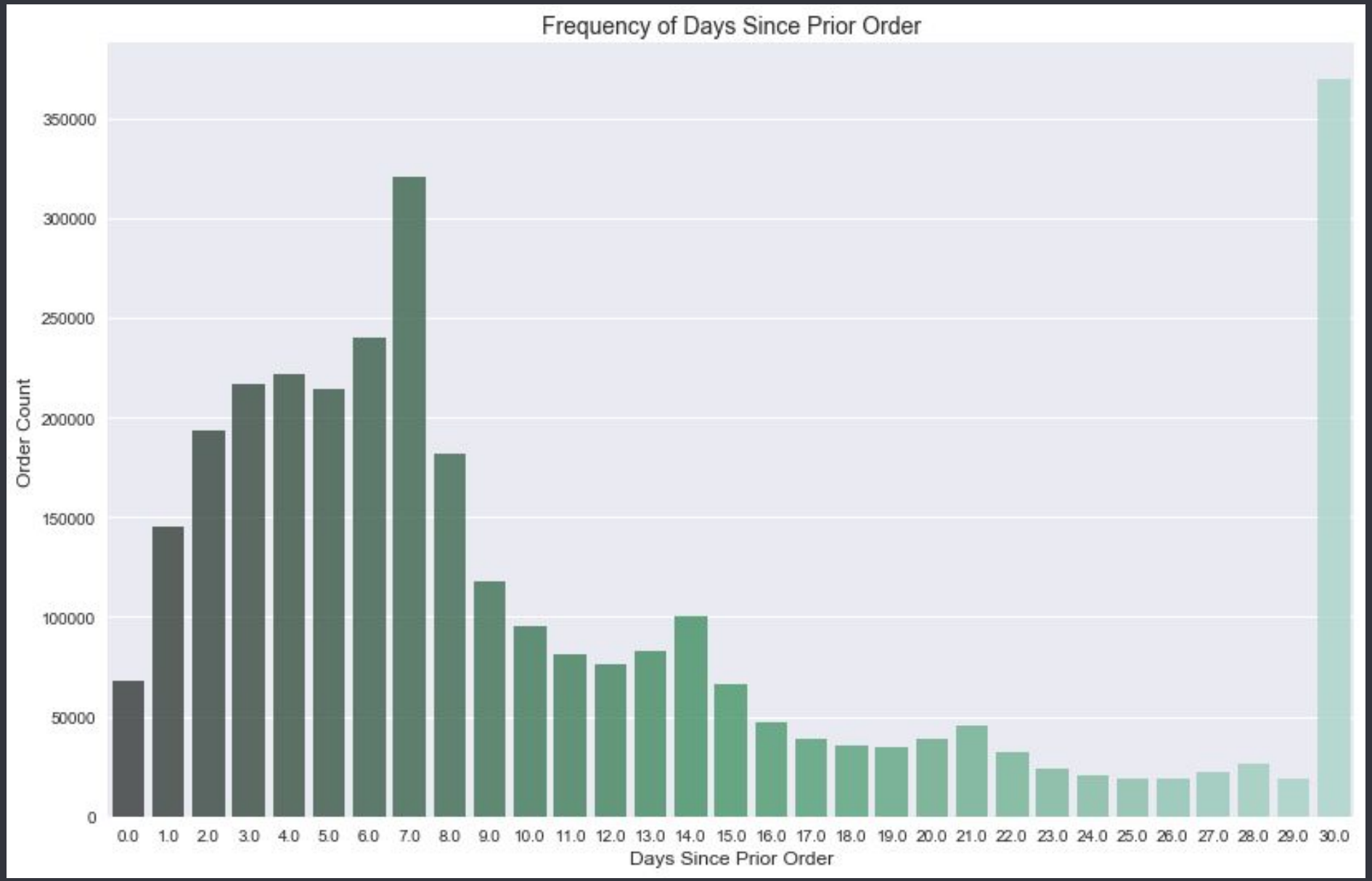
Data consists of over 3 million orders from more than 200K users.

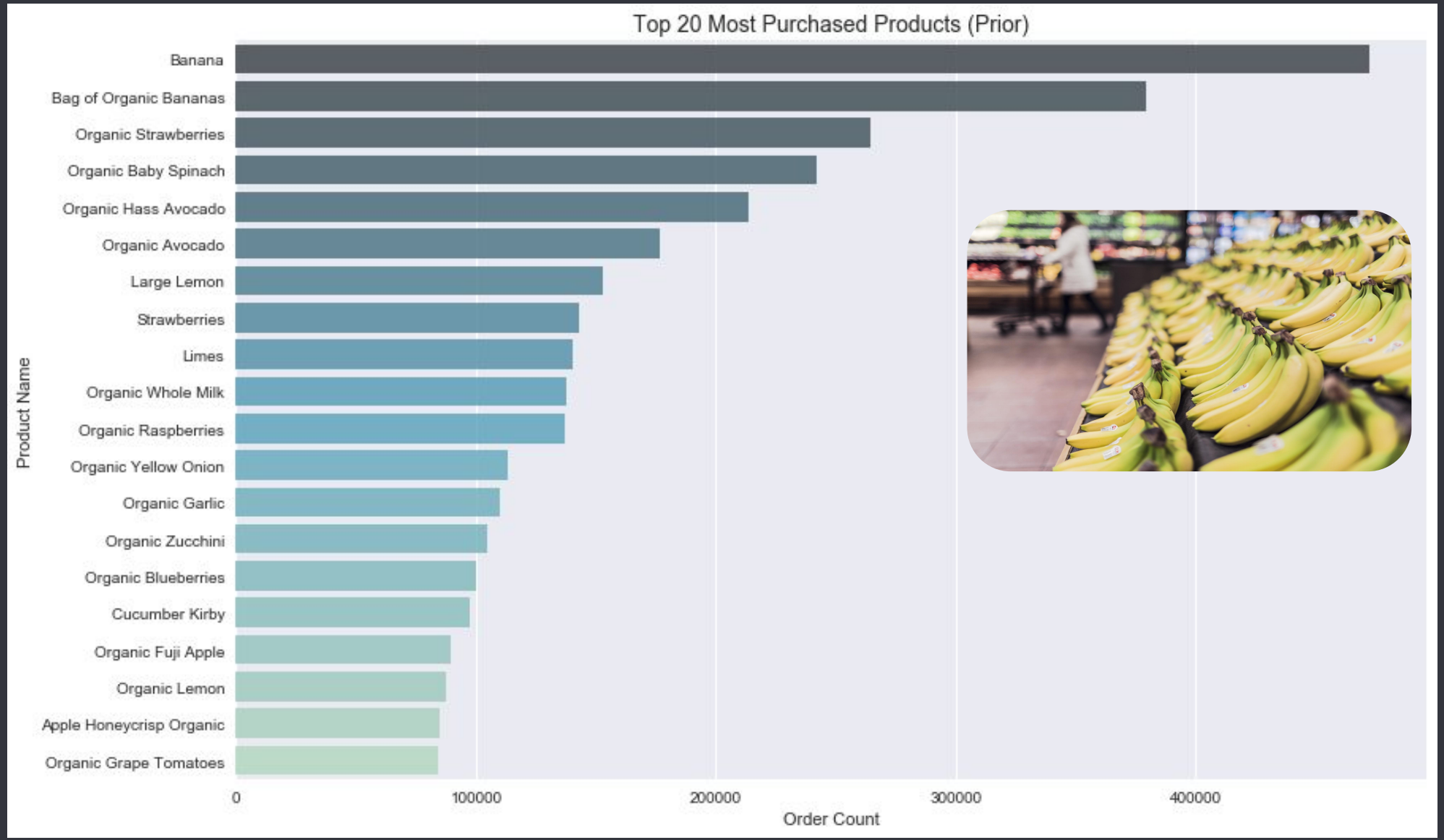
Available information: user order count, order products were added to a user's cart, day of week, hour of day, reordered indicator, product department, product aisle.


Unique Customers
Users
200000
175000
150000
125000
100000
75000
50000
25000
0
prior
test
train
Eval Set


Top 30 Orders User Count
User Count
25000
20000
15000
10000
5000
0
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Max Orders


Frequency of Day of week Vs Hour of day
Day of Week
0
1
2
3
4
5
6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day
50000
40000
30000
20000
10000


Frequency of Days Since Prior Order
Order Count
0
50000
100000
150000
200000
250000
300000
350000
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0 26.0 27.0 28.0 29.0 30.0
Days Since Prior Order


Top 20 Most Purchased Products (Prior)
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Large Lemon
Strawberries
Limes
Organic Whole Milk
Organic Raspberries
Organic Yellow Onion
Organic Garlic
Organic Zucchini
Organic Blueberries
Cucumber Kirby
Organic Fuji Apple
Organic Lemon
Apple Honeycrisp Organic
Organic Grape Tomatoes
Product Name
0
100000
200000
300000
400000
Order Count

Default Variable Correlation

	order_number	order_dow	order_hour_of_day	days_since_prior_order	product_id	add_to_cart_order	aisle_id	department_id
order_number	1.0	0.016	-0.04	-0.36	-0.0021	-0.0037	0.0051	0.0042
order_dow	0.016	1.0	0.014	-0.03	-0.0058	-0.0087	0.00057	0.0062
order_hour_of_day	-0.04	0.014	1.0	0.0017	0.00061	-0.015	-0.0036	-0.0097
days_since_prior_order	-0.36	-0.03	0.0017	1.0	0.002	0.053	0.0051	-0.00013
product_id	-0.0021	-0.0058	0.00061	0.002	1.0	0.0056	0.0033	-0.028
add_to_cart_order	-0.0037	-0.0087	-0.015	0.053	0.0056	1.0	0.0076	0.028
aisle_id	0.0051	0.00057	-0.0036	0.0051	0.0033	0.0076	1.0	0.063
department_id	0.0042	0.0062	-0.0097	-0.00013	-0.028	0.028	0.063	1.0

Highly correlated variables:

'order_dow' and 'order_hour_of_day'

'add_to_cart_order' and 'days_since_prior_order'

Feature Selection

Developed 13 features to evaluate for predictive performance using Random Forest Classifier.

	features	importance
16	user_prod_reorder_rate	0.4873
19	prod_reorder_rate	0.0584
8	user_orders	0.0517
15	avg_add_to_cart_order	0.0435
11	avg_reorders_per_basket	0.0358

Selected Features

User prod reorder rate: user frequency of reordering product

Prod reorder rate: frequency product is reordered (by all users)

Avg add to cart order: average sequence a product is added to a user's order

Avg reorders per basket: proportion of reorded products in a user's orders

63.0%

Products reordered in the training (prior order) data set.

Products were reordered in the validation (test order) data set.

59.8%

Baseline performance

(Using sklearn LR evaluation)

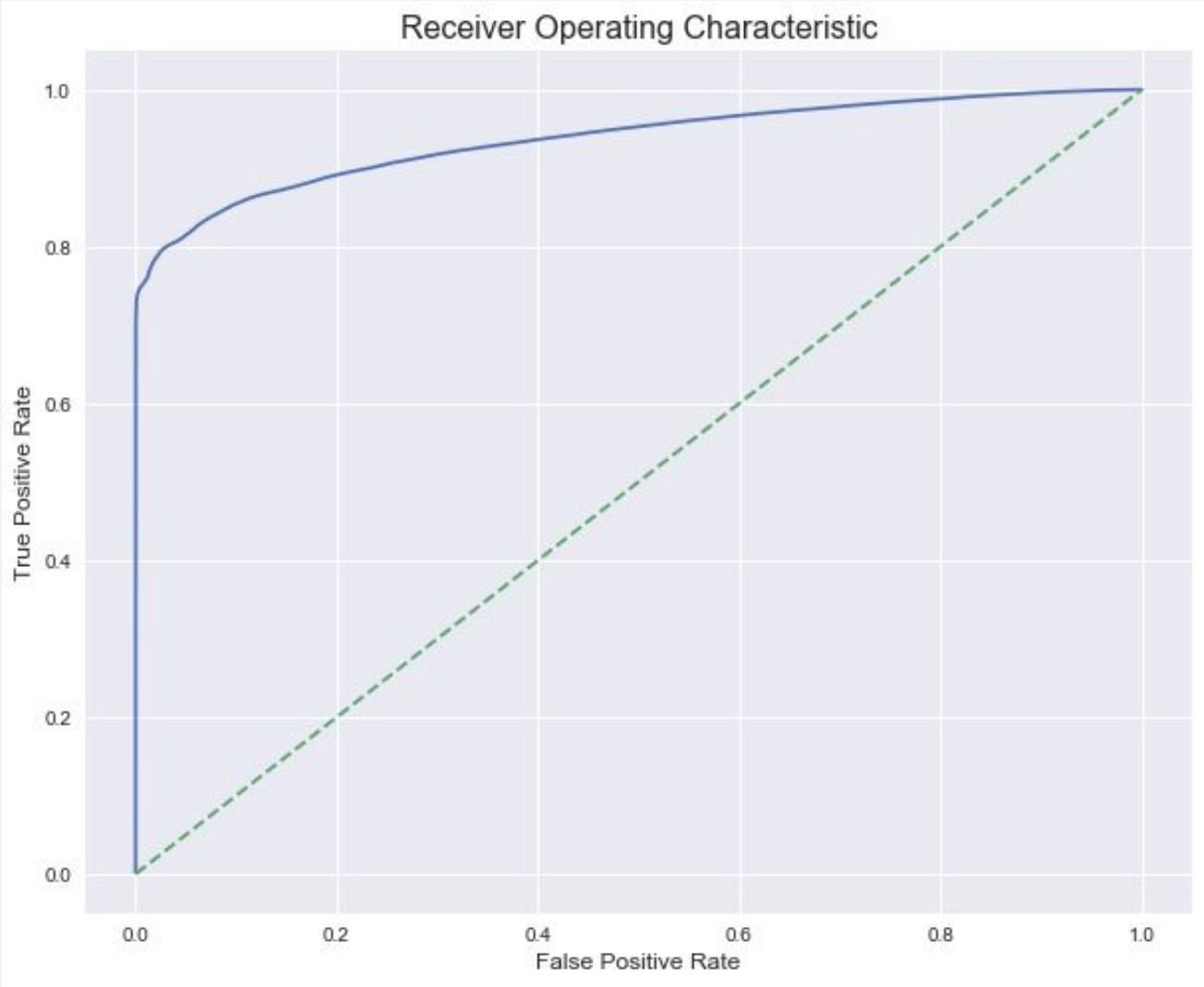
Accuracy 77.7%

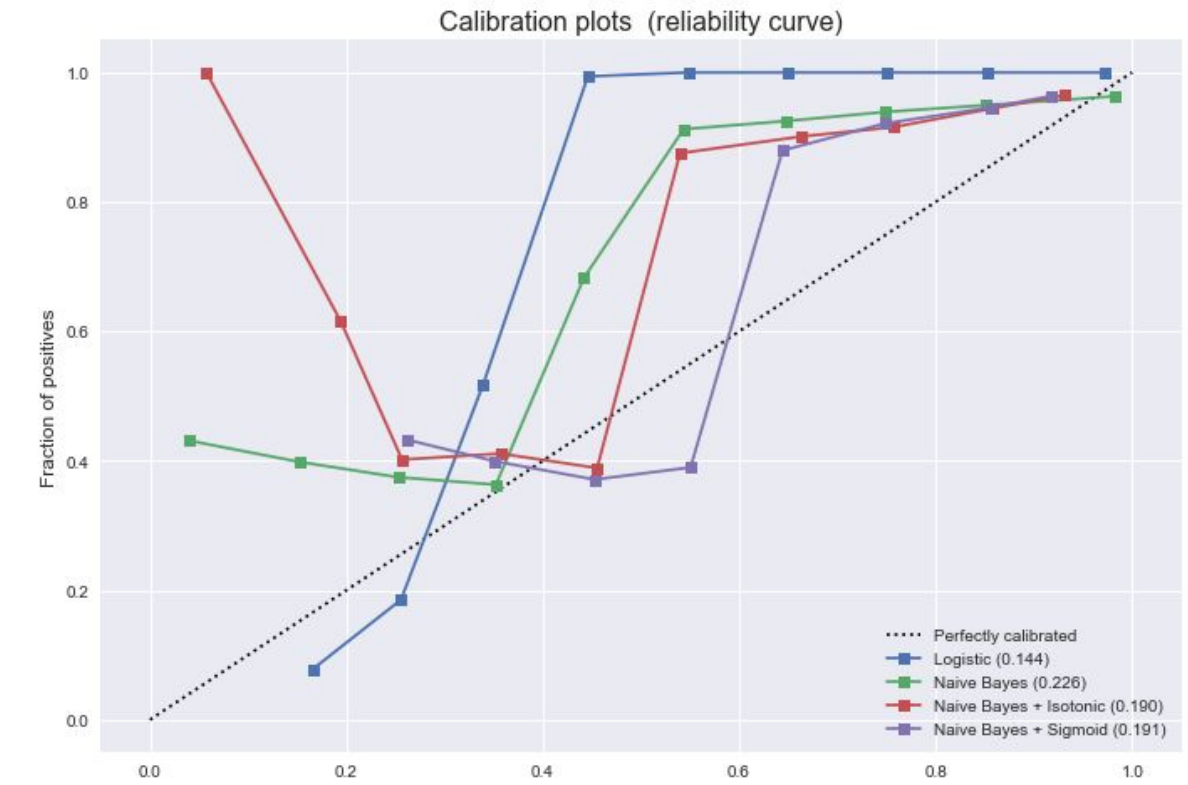
Precision 100%

Recall 62.7%

F1 0.771

AUROC 93.7%





SKLearn Calibration Plots

Native Bayes + Sigmoid demonstrates the best reduction in overfitting.

Logistic:
- Brier: 0.144
- Precision: 1.000
- Recall: 0.627
- F1: 0.771

Naive Bayes + Isotonic:
- Brier: 0.190
- Precision: 0.932
- Recall: 0.583
- F1: 0.717

Naive Bayes:
- Brier: 0.226
- Precision: 0.945
- Recall: 0.486
- F1: 0.642

Naive Bayes + Sigmoid:
- Brier: 0.191
- Precision: 0.738
- Recall: 0.727
- F1: 0.733

SKLearn Calibration Plots

SVC + Isotonic demonstrates the best overall performance.

Logistic:

- Brier: 0.144
- Precision: 1.000
- Recall: 0.627
- F1: 0.771

SVC + Isotonic:

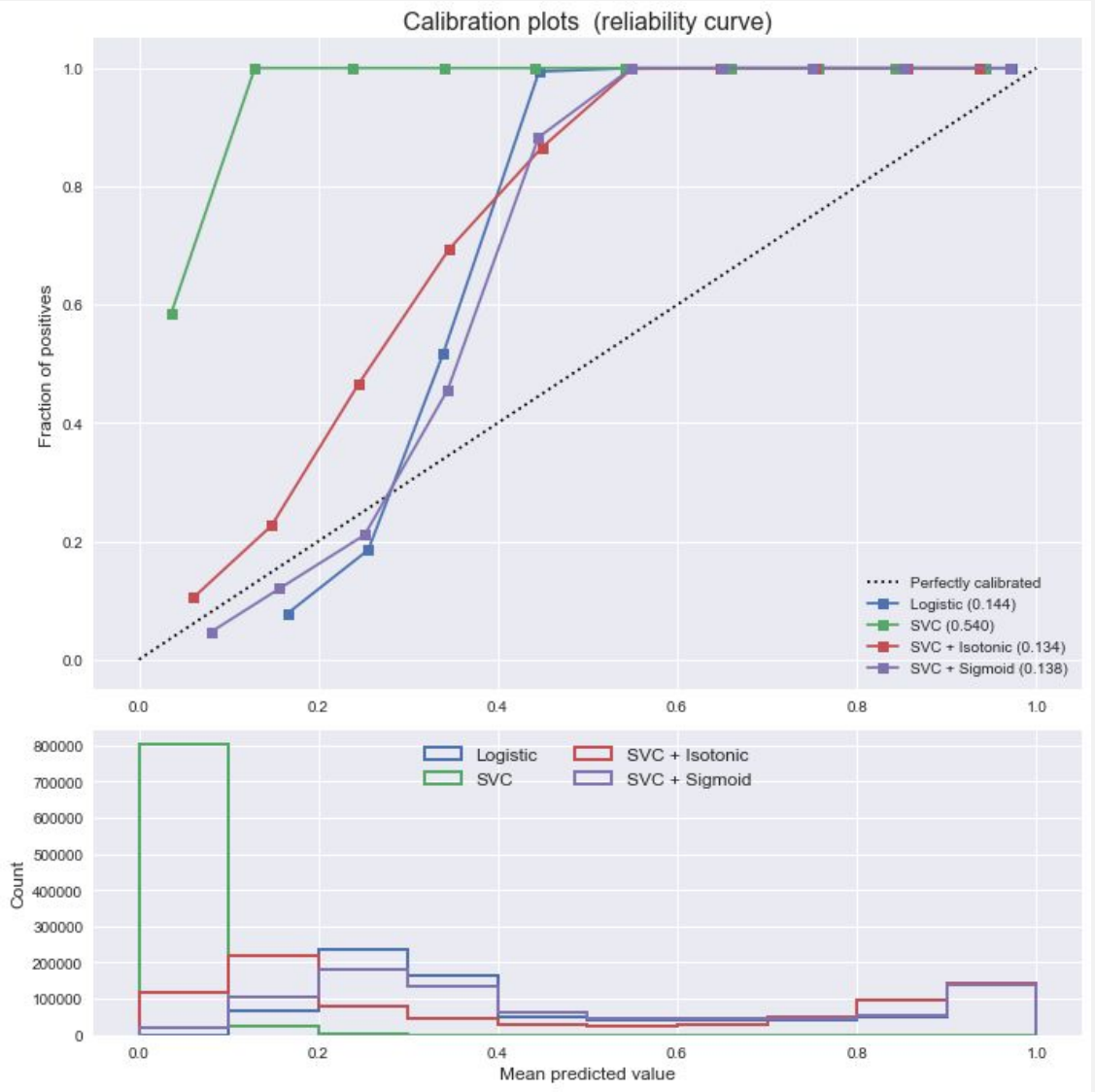
- Brier: 0.134
- Precision: 1.000
- Recall: 0.690
- F1: 0.817

SVC:

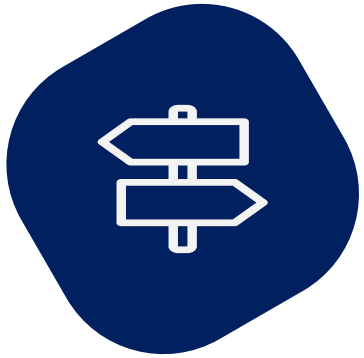
- Brier: 0.540
- Precision: 1.000
- Recall: 0.659
- F1: 0.794

SVC + Sigmoid:

- Brier: 0.138
- Precision: 1.000
- Recall: 0.658
- F1: 0.794

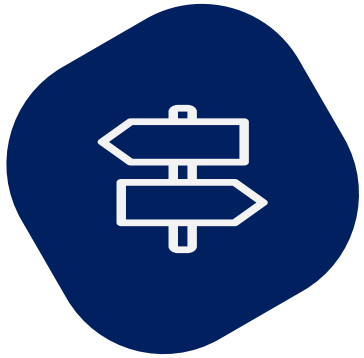


Reflections



- ➔ Developing a well-balanced model is difficult with only a single highly predictive feature.
- ➔ While performance appears promising when applied to customers' next order, this can not be necessarily anticipated with future orders.
- ➔ The combined linear support-vector classifier (LinearSVC) + Isotonic model had the best overall performance. It is less than ideal, evidenced by its distance from isotonic regression line (diagonal). This is likely caused by the dominance of a single feature in the model.

Reflections



Additional questions to explore:

- Which product is a customer likely to try for the first time during their next order?
- When will a customer make their next order?
- What customer segments can be derived from purchasing behavior?
- What products are commonly purchased together?

Thank You

Sean Matthews

github/sean-io/market-basket-analysis